Statistical modeling of spatial big data: An approach from a functional data analysis perspective

Ramón Giraldo [a,*], Sophie Dabo-Niang [b], Sergio Martínez [a]

[a] *Statistics Department, Universidad Nacional de Colombia, Colombia*
[b] *Université Lille 3, Laboratoire LEM CNRS, France*

A literature review on spatial big data analysis is given. We show an application of Universal Kriging to a massive spatial dataset. We also present some perspectives of future work in this field.

1. Introduction

Modern technology has facilitated the collection of very large datasets. Nowadays, we generate about several trillion bytes of data every day, characterized by high dimensionality and large sample size and called Big Data or massive volumes of data. The term Big data is relatively new, but gathering and storing huge amounts of data is an old subject. The early years of the new millennium saw the concept of large dataset (Laney, 2011). However, due to its complexity, no universal definition can be given to big data.

The last two decades have seen an extensive development of statistical tools capable of managing huge quantities of data consisting of a matrix of n data vectors, each containing p measurements, usually with large n and p (Izenman, 2008). Among others Functional Data Analysis (FDA) (Ramsay and Silverman, 2005) has arisen as a field of statistics for modeling data in this context, particularly when $p \gg n$.

FDA is concerned with the modeling of data, such as curves, shapes, images or a more complex mathematical object, thought as smooth realizations of a stochastic process (which is an infinite dimensional data object). In this sense functional data are part of Big Data (Zipunnikov et al., 2011; Chen et al., 2011, 2015). It is common to find applications of FDA where the size of data is of order of terabytes, as encountered for example in fMRI (functional magnetic resonance imaging), brain imaging analysis (Chen et al., 2015) or bioinformatics (Yoo et al., 2014).

In a number of diverse disciplines such as environmental sciences, agronomy or mining, the data have an inherent spatial component. Spatial statistics (Cressie and Wikle, 2011) embodies a suite of methods for analyzing this type of information. This characteristic is also present in some spatial massive dataset (called spatial big data). An example is when long time series of meteorological variables are recorded at each point of a monitoring network or space–time usage of the mobile-phone network in a given area (Secchi et al., 2015).

In this contribution, we highlight the use of FDA to model spatial big data. We briefly review in Section 2 the notion of spatial big data and some FDA tools that can be used to model such data, particularly for doing spatial prediction. Section 3 is dedicated to an illustration on a big spatial dataset of daily temperature curves. The article ends with a brief discussion and suggestions for further research in Section 4.

* Corresponding author.
E-mail address: rgiraldoh@unal.edu.co (R. Giraldo).

2. A review on modeling spatial big data by using functional data analysis

Today statistical techniques for the analysis of large and complex data with a spatial underlying structure are required. For example, it is the case in neurological studies when curves of the electrical activity are recorded in voxels of the brain (Lindquist, 2008). Generally, some multivariate techniques of dimension reduction are used to solve the problem of high dimensionality that arises in these cases (Izenman, 2008). FDA may be an alternative in this scenario. High dimensional data become functional data which are posteriorly analyzed by using FDA methods (Giraldo et al., 2011). Basically, it is considered that the data observed for several dependent spatial units correspond to realizations of stochastic processes, usually assumed to be in a metric, semi-metric, Hilbert or Banach space.

In fact, in the setting of functional spatial data analysis, complex high dimensional data is typically a set of curves or surfaces, spatially distributed, regarded as points in a functional space (space of squared integrable functions, …).

Here, we give a review about the common basis of these approaches making an emphasis in the spatial prediction of functional data.

The last decade saw a dynamic literature on parametric or non-parametric FDA approaches for high dimensional spatially dependent data and applications to various domains, such as principal component analysis (Li and Guan, 2014; Liu et al., 2017), clustering (Giraldo et al., 2012; Secchi et al., 2013; Romano et al., 2015), regression and prediction of stationary or non-stationary processes with kriging or non-parametric methods (Ignaccolo et al., 2008; Dabo-Niang et al., 2010; Nerini et al., 2010; Delicado et al., 2010; Giraldo et al., 2011; Dabo-Niang et al., 2012; Menafoglio and Secchi, 2013; Sangalli et al., 2013; Caballero et al., 2013; Ternynck, 2014; Zhu et al., 2014; Bohorquez et al., 2016, 2017), testing (Aston et al., 2017), among others. For more details, see the recent review of Menafoglio and Secchi (2017) on object oriented spatial statistics and references therein.

In the following, we propose to illustrate spatial prediction involving complex high dimension datum by a modern functional prediction technique, namely, functional kriging.

3. Spatial prediction of temperature curves

In this section we show briefly an application of universal kriging for functional data (Caballero et al., 2013) as an example of a methodology for modeling spatial big data. We consider daily temperature data recorded at 772 stations from the meteorological monitoring network of Colombia (Fig. 1). Specifically, we have 13.149 data at each station corresponding to daily records of maximum temperature (°C) obtained from January 1, 1980 to December 31, 2015. Note that this is a spatiotemporal dataset and could be analyzed by using, among other methodologies, space–time geostatistics (Christakos, 2000). However, given the high volume of data it is reasonable or even necessary to assume an approach based on FDA.

In Fig. 1 we show the temperature data of a sample of ten stations (of the total of 772) randomly selected. We only show the data in this small sample of stations in order to identify more easily the temporal behavior of the temperature. In a first step of the analysis the discrete data at each station are converted into curves by using smoothing methods. Specifically, we expand each time series in terms of 100 Fourier basis functions. The number of basis functions was chosen by using cross-validation. The temperature in Colombia varies considerably according to the topography. For this reason we take the altitude as a covariable into the analysis. As an example, we carry out prediction on an unvisited location (red point in Fig. 1). In order to reduce the computational time, we use as input the curves corresponding to the 30 closest stations to the prediction site. The smoothed curves in these sites and the prediction at the unvisited site considered are shown in Fig. 2.

The methodology here illustrated has been replicated for the Colombian Institute of Hydrology, Meteorology and Environmental Studies as a procedure for getting information of this variable in remote and difficult to access zones of Colombia. This technique has been also used to predict the temperature values in large time periods with missing values.

4. Further research

The purpose of the present work shows the relevancy of using the functional framework in analyzing massive spatial data. After an introduction to functional data as part of big data and a review, a spatial prediction method is used to forecast high dimension spatial temperature data.

Conceptually, FDA is suited for spatial huge data as shown by the illustration presented in this paper. However, this last is relatively simple and deals with a single functional object. In fact, spatial big data will become increasingly common, and a large number of complex situations with many potential applications may be considered. The statistical community must then address the challenge of proposing new methods for describing and analyzing spatial big data with complex structures. For instance, one may consider, different big and non-big data objects of different types (qualitative, quantitative, ordinal, …) for different regions and pay attention to hot topics such as, missing (for instance a continuous part of a curve) or extreme data. Particularly, in the meteorological context considered in Section 3, the modeling of satellite information (an example of massive spatial datasets) in combination with ground data or taking into account the peak (maximum) in the smoothing procedure, is an open research field. For FDA, reaching the peaks necessitates an important amount of basis functions. In addition, setting a FDA approach requires a proper definition of the targeted function spaces, suitable metrics to measure similarity between objects, spatial or space–time correlations techniques, ….

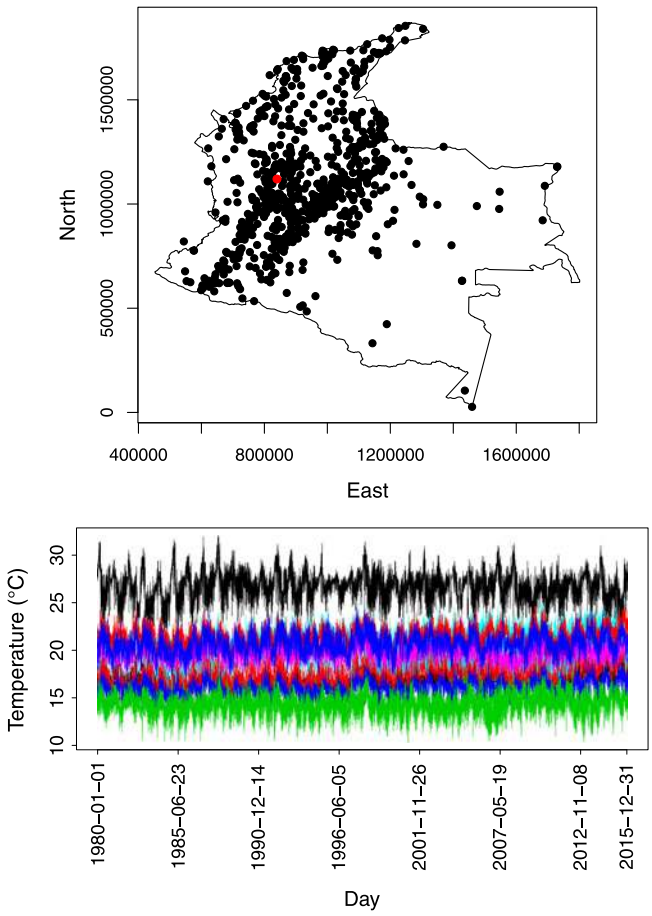

Fig. 1. Top: Temperature monitoring network of Colombia (black points). Red point corresponds to a station without information. Bottom: Records of daily maximum temperature (°C) obtained from January 1, 1980 to December 31, 2015 in ten stations (randomly chosen) from the Colombian meteorological monitoring network. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

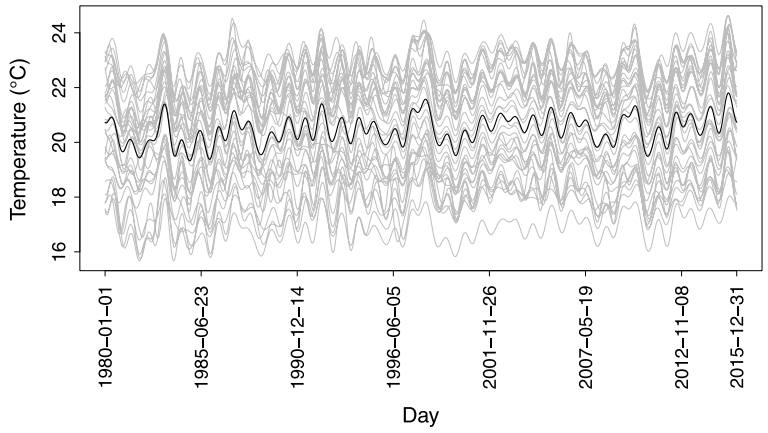

Fig. 2. Temperature curves obtained by smoothing the data with a Fourier basis (gray curves). Prediction of a smoothed curve on an unvisited station (black curve).

In fact, a main challenge when analyzing (dimension reduction, clustering, parametric, non-parametric, semi-parametric, additive regression models, analysis of variance, tests, ...) complex spatial big data is to use statistical tools able to compute in a non-costly way (Cressie and Wikle, 2011) spatial correlations among huge amounts of data (Zipunnikov et al., 2011; Chen et al., 2015).

References

Aston, J.A., Pigoli, D., Tavakoli, S., et al., 2017. Tests for separability in nonparametric covariance operators of random surfaces. Ann. Statist. 45 (4), 1431–1461.

Bohorquez, M., Giraldo, R., Mateu, J., 2016. Optimal sampling for spatial prediction of functional data. Stat. Methods Appl. 25 (1), 39–54.

Bohorquez, M., Giraldo, R., Mateu, J., 2017. Multivariate functional random fields: prediction and optimal sampling. Stoch. Environ. Res. Risk Assess. 31 (1), 53–70.

Caballero, W., Giraldo, R., Mateu, J., 2013. A universal kriging approach for spatial functional data. Stoch. Environ. Res. Risk Assess. 27 (7), 1553–1563.

Chen, K., Chen, K., Müller, H.-G., Wang, J.-L., 2011. Stringing high-dimensional data for functional analysis. J. Amer. Statist. Assoc. 106 (493), 275–284.

Chen, K., Zhang, X., Petersen, A., Müller, H.-G., Wang, J.-L., 2015. Quantifying infinite-dimensional data: functional data analysis in action. Stat. Biosci. 1–23.

Christakos, A., 2000. Modern Spatiotemporal Geostatistics. Oxford University Press, Oxford.

Cressie, N., Wikle, C.K., 2011. Statistics for Spatio-Temporal Data. In: Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., Hoboken, NJ.

Dabo-Niang, S., Kaid, Z., Laksaci, A., 2012. On spatial conditional mode estimation for a functional regressor. Statist. Probab. Lett. 82 (7), 1413–1421.

Dabo-Niang, S., Yao, A.-F., Pischedda, L., Cuny, P., Gilbert, F., 2010. Spatial mode estimation for functional random fields with application to bioturbation problem. Stoch. Environ. Res. Risk Assess. 24 (4), 487–497.

Delicado, P., Giraldo, R., Comas, C., Mateu, J., 2010. Statistics for spatial functional data: some recent contributions. Environmetrics 21 (3–4).

Giraldo, R., Delicado, P., Mateu, J., 2011. Ordinary kriging for function-valued spatial data. Environ. Ecol. Stat. 18 (3), 411–426.

Giraldo, R., Delicado, P., Mateu, J., 2012. Hierarchical clustering of spatially correlated functional data. Stat. Neerl. 66, 403–421.

Ignaccolo, R., Ghigo, S., Giovenali, E., 2008. Analysis of air quality monitoring networks by functional clustering. Environmetrics 19 (7), 672–686.

Izenman, A.J., 2008. Modern Multivariate Statistical Techniques. In: Springer Texts in Statistics, Springer, New York, regression, classification, and manifold learning.

Laney, D., 2011. 3D Data management: controlling data volume, velocity, and variety. META Group.

Li, Y., Guan, Y., 2014. Functional principal component analysis of spatiotemporal point processes with applications in disease surveillance. J. Amer. Statist. Assoc. 109 (507), 1205–1215.

Lindquist, A., 2008. The statistical analysis of fMRI data. Statist. Sci. 23 (4), 439–464.

Liu, C., Ray, S., Hooker, G., 2017. Functional principal component analysis of spatially correlated data. Stat. Comput. 27 (6), 1639–1654.

Menafoglio, A., Secchi, P., 2013. A universal kriging predictor for spatially dependent functional data of a Hilbert space. Electron. J. Stat. 7, 2209–2240.

Menafoglio, A., Secchi, P., 2017. Statistical analysis of complex and spatially dependent data: a review of Object Oriented Spatial Statistics. European J. Oper. Res. 258 (2), 401–410.

Nerini, D., Monestiez, P., Manté, C., 2010. Cokriging for spatial functional data. J. Multivariate Anal. 101 (2), 409–418.

Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis, second ed. In: Springer Series in Statistics, Springer, New York.

Romano, E., Mateu, J., Giraldo, R., 2015. On the performance of two clustering methods for spatial functional data. Adv. Stat. Anal. 99 (4), 467–492.

Sangalli, L.M., Ramsay, J.O., Ramsay, T.O., 2013. Spatial spline regression models. J. R. Stat. Soc. Ser. B Stat. Methodol. 75 (4), 681–703.

Secchi, P., Vantini, S., Vitelli, V., 2013. Bagging voronoi classifiers for clustering spatial functional data. Int. J. Appl. Earth Obs. Geoinf. 22, 53–64.

Secchi, P., Vantini, S., Vitelli, V., 2015. Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan. Stat. Methods Appl. 24 (2), 279–300.

Ternynck, C., 2014. Spatial regression estimation for functional data with spatial dependency. J. Soc. Franc. Statist. 155 (2), 138–160.

Yoo, C., Ramirez, L., Liuzzi, J., 2014. Big data analysis using modern statistical and machine learning methods in medicine. Int. Neurourol. J. 18, 50–57.

Zhu, H., Fan, J., Kong, L., 2014. Spatially varying coefficient model for neuroimaging data with jump discontinuities. J. Amer. Statist. Assoc. 109 (507), 1084–1098.

Zipunnikov, V., Caffo, B., Yousem, D.M., Davatzikos, C., Schwartz, B.S., Crainiceanu, C., 2011. Functional principal component model for high-dimensional brain imaging. NeuroImage 58 (3), 772–784.